A COMPARISON OF MULTIPLE REGRESSION AND A NEURAL NETWORK FOR PREDICTING A MEDICAL DIAGNOSIS

W. M. Pugh

Report No. 91-33

NAVAL HEALTH RESEARCH CENTER
P.O. BOX 85122
SAN DIEGO, CALIFORNIA 92186-5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

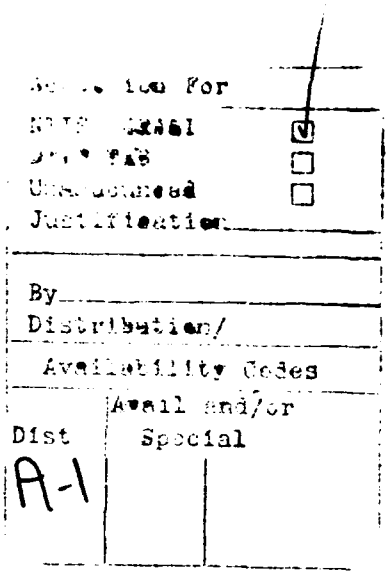
A COMPARISON OF MULTIPLE REGRESSION AND A NEURAL NETWORK FOR PREDICTING A MEDICAL DIAGNOSIS

William M. Pugh

Naval Health Research Center

P.O. Box 85122

San Diego, CA 92186-5122



Report No. 91-33 was supported by the Naval Military Research and Development Command, Department of the Navy, Work Unit No. 63706N-M0095.005-6051. The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor any other department or agency of the United States Government.

SUMMARY

Problem

A variety of methods have been used to develop algorithms for medical diagnosis. Two methods that have been used recently are multiple regression and neural networks. The current study is a systematic comparison of these two methods using simulated data.

Approach

Data sets were developed using four variables which were combined in three ways. One combination was a linear function, the second was a polynomial function, and the third function included an interactive term. For each function, a small data set of 100 cases, and a large data set of 1,000 cases was generated.

The methods were evaluated by developing predictive algorithms using a regression program and a neural network program. Validation correlation coefficients were produced by correlating the predicted score from each method with true scores on the data set used to develop the algorithm. The predictive algorithm was also applied to another data set and these values were correlated with the true scores to generate cross-validation correlations.

Results

Results of the analyses for the data combined using a linear function reveal virtually no difference between the regression and neural network methods for large samples. On the small data sets the neural network algorithms tended to overfit on the

validation samples and displayed a relatively large degree of shrinkage upon cross validation.

Analysis of the data sets containing a polynomial term or an interactive term showed that the neural network technique could fit the non-linear variance without any specification being provided. The regression program required the form of the non-linear term to be specified before the regression technique would fit it. These analyses also showed that the neural network had a greater tendency than the regression program to overfit when provided small data sets, and as a result, the regression equation generalized better on such data.

Conclusions

It was concluded that neural networks could be useful when developing diagnostic algorithms by locating non-linear effects. However, such analyses would require data sets sufficiently large to avoid fitting sample-specific variance. In this regard it is recommended that a cross-validation procedure be used and the degree of shrinkage between the validation and cross-validation samples be examined. A large degree of shrinkage would indicate the algorithm would not generalize to other samples. Finally, it was suggested that in the event that a non-linear relationship is found, the investigator should attempt to specify the form of the relationship and use that information to build a regression equation.

A COMPARISON OF MULTIPLE REGRESSION AND A NEURAL NETWORK FOR PREDICTING A MEDICAL DIAGNOSIS

William M. Pugh and David H. Ryman

The increased availability of computer technology has been paralleled by a series of efforts to apply analytical methods to medical decision making. Early work (Ledley and Lusted, 1959) involved the application of Bayes Theorem to the task of deriving the probability of a medical diagnosis. Subsequent investigators have applied a variety of techniques including multiple regression, linear discriminant analysis (Titterington, Murray, G. D., Murray, L. S., Spiegelhalter, D. J., 1981), and expert systems (Kinney, Brafman, Wright, 1988).

Navy researchers have been particularly interested in these developments because of the potential application to small U.S. Navy ships and submarines. The medical department on these vessels is typically run by a specially trained independent duty hospital corpsman. If one of these corpsmen is confronted with a patient suffering from an acute illness that is difficult to diagnose, the ship may have to abort its mission so that medical consultation and support can be obtained. Therefore, work was begun on the development of diagnostic algorithms to provide support for corpsmen at sea (Stetson, Eberhart, Dobbins, Pugh, Gino, 1990).

The initial algorithms developed for Navy personnel used either a rule-based approach (Newacheck, 1990) or a bayesian approach (de Dombal, Leaper, Staniland, McCann & Horrocks, 1972; Carras, Southerland, Fisherkeller, 1989). More recently the utility of using neural network technology for medical diagnosis has been investigated (Eberhart & Dobbins, 1989). The current study was undertaken to better understand the strengths and weaknesses of the neural network approach by comparing that approach to the more established analytic technique of multiple regression. This comparison was accomplished using a computer simulation in which randomly generated signs and symptoms were combined according to three different models to create the diagnostic outcome. Applying the neural network and multiple regression procedures to these data with known characteristics allowed the two techniques to be compared in a systematic fashion.

Method

Techniques

The regression analyses were conducted with the SPSSX multiple regression program. The neural network used was the Batchnet program developed by the Eberhart and Dobbins (1990). The Batchnet program is a feedback learning routine with options that allows the user to specify the number of hidden layers, the

number of hidden nodes, learning rate, and activation function. All data points, however, must range between 0 and 1.

Data

Predictor variables. Thirty data sets of 100 simulated patients and 10 data sets of 1000 simulated cases were generated. Four predictor variables, i.e., data representing medical signs and symptoms, were created using the SPSSX random number generator. Although many symptoms are normally distributed others are not. To address this issue while attempting to limit the complexity of the simulation it was decided that two common distributions would be used in generating the predictor variables. Therefore, the first two variables, X_1 and X_2, were created as normally distributed variables with a mean of 0.5 and a standard deviation of 0.3. The second two variables, X_3 and X_4, were uniformly distributed between 0 and 1.

Outcome variables. Three different outcome or criterion scores were created to reflect different ways that signs and symptoms could be related to a medical diagnosis. In each situation, three separate effects are combined to produce the outcome measure. The first type of relationship is one where the three conditions contributed equally to the outcome, and each additional increase in a predictor measure produces a commensurate increase in the manifestation of the medical disorder. For example, fever, abdominal pain, and nausea may all

lead to a diagnosis of appendicitis. Further, if an increase in any or all of these may indicate a greater manifestation of the disorder, the situation may be modeled as follows:

$$Y_1 = \frac{X_1 + X_2 + X_3 + \sqrt{2}E}{3 + \sqrt{2}}$$

This function will be referred to as the linear function.

An error term (E) was added to the signs and symptoms (X_1, X_2 and X_3) to represent the uncontrolled variables and errors of measurement that prevent one from being certain about every diagnosis. E was created as a normally distributed random variable with a mean of 0.5 and a standard deviation of 0.3. The result was weighted by the square root of two so that it would account for 40 percent of the variance in the outcome Y_1. Such an amount of error seemed to be a plausable estimate of the variance resulting from errors of measurement and unmeasured symptoms or causal factors. But, more importantly it was felt that this would test the ability of the two techniques to separate the signal (predictable variance) from the noise (error). Dividing by $3+\sqrt{2}$ simply brought the Y_1 values into 0 to 1 range required for Batchnet.

The second type of relationship that was examined both represented a situation where both a low and a high value on a predictor variable was indicative of the diagnosis. For example, loss of consciousness, low-blood pressure, and either too little

or too much insulin, could be indicative of diabetic shock. The effect of insulin, in this case, can be represented by a second degree polynomial S which is a function of X_3.

To keep S within the range 0 to 1, S was created as follows:

$$S=4(X_3-.5)^2$$

Then, the above situation was modeled as follows:

$$y_2=\frac{X_1+X_2+S+\sqrt{2}E}{3+\sqrt{2}}$$

This function will be called the polynomial function.

The third type of relationship considered is one where two variables interact. An example of this type of relationship would be seen with the diagnosis of heart disease. In addition to risk factors such as excessive weight and hypertension, factors such as diet and cholesterol levels could be interactive. For instance, a fatty diet and low cholesterol levels could indicate a patient has the ability to metabolize cholesterol very well while a low fat high fiber diet with high cholesterol levels may suggest a genetic propensity toward high cholesterol levels.

This relationship of diet and cholesterol level can be xpressed as (I) the product of the measures (X_3 and X_4). To eep I within the range of 0 to 1, I was created as follows:

$$I=2(X_3-.5)(X_4-.5)$$

The above relationship was, then, modeled as follows:

$$Y_3=\frac{X_1+X_2+I+\sqrt{2}E}{3+\sqrt{2}}$$

nd will be called the interactive function.

esign

Analyses were conducted on thirty data sets of 100 cases, nd ten data sets of 1000 cases. Information from each data set as entered into the multiple regression and the neural network rograms. The predictor variables were organized into two sets, ne containing only the linear variables X_1, X_2, X_3, and X_4, and he second set consisted of these linear variable plus the two on-linear terms S and I. Each predictor set was used to develop he prediction parameters employed by the respective analytic echniques for each of the outcome variables (Y_1, Y_2, and Y_3). hus, for each data set 12 predictions were made: an algorithm

as derived from each of the two predictor sets, for each of the hree outcome variables, from each of the two techniques.

rocedures

A series of pilot analyses conducted prior to this study ndicated that for data sets such as the ones used in the current nalyses, the neural network performed best when one hidden layer ith 5 nodes and a learning rate of 0.15 was used. There these arameters were used along with a sigmoid activation function, .e.,

$$F(x)=\frac{1}{1+e^{-x}}$$

nce an analysis was run on a data set, the parameters produced ere applied to that data set to yield a predicted value for each utcome variable (i.e., $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$). The accuracy of these redictions was evaluated by correlating them with the actual y alues. These correlations ($r_{y\hat{y}}$) computed on the data set used o derive the prediction parameters are referred to as the alidation correlations.

To assess the degree that these prediction methods would eneralize, the parameters from each data set were applied to the redictor variables of the next data set. For instance, the arameters of data set one were applied to the predictor

ıriables in data set two, producing predicted y values for data
:t two. Parameters from the last data set were applied to the
.rst one. These $\hat{y}$ values generated using parameters derived on
ıother data set were correlated with the actual y values to
'oduce a set of correlations referred to as the cross-validation
›rrelations.

Results

Inspection of the predictor data showed that each of the
'edictor terms had approximately the same standard deviation
:cept for I, which had a standard deviation approximately half
ıe size of the others. This result can be used to compute the
›rrelations that would be obtain with an optimal predictor
.e., a predictor of everything but the error term). This is
›ne by letting the standard deviation for each predictor term be
except for the standard deviation of I which would be 1/2 D.
so, the standard deviation of E would be equal to D. Because
l the terms were independently generated, the covariances can
assumed to be zero, and the total variance for any function
n be found by summing the component variances. Thus, the
riance contained in the linear function (Y_1) can be found by
mming the variance of X_1, X_2, and X_3, and then adding the error
riance as follows:

$$3D^2+(\sqrt{2}D)^2=5D^2$$

Subtracting the amount of error variance from the total leaves the amount of predictable variance; $3D^2$. The ratio of the predictable to total variance is the correlation ratio:

$$CR = \frac{3D^2}{5D^2}$$

and the square root of the correlation ratio yields the correlation ($r = .77$) that would be obtained with the optimal predictor. In a similar fashion, correlation ratios can be computed for the polynomial and interactive functions. In addition, the numerator can be decomposed into the linear and non-linear components. Table 1 shows the results of carrying out these computations. The first column of values are the validation correlations that would be expected if the linear variance was correctly predicted, and the second column are the validation correlations that would be expected if all predictable variance (both linear and non-linear) was accounted for.

Table 1

Expected Correlations for Optimal Prediction Parameters

	Correlation	
Function	Linear Variance	Predictable Variance
Linear	.77	.77
Polynomial	.63	.77
Interactive	.69	.73

Linear Function

The results of data analyses conducted with respect to the linear function are shown in Table 2.

Table 2

Mean Correlations Between Actual and Predicted Criterion Scores Generated Using a Linear Function

		PREDICTOR SET			
NO. CASES	TECHNIQUE	LINEAR TERMS ONLY		LINEAR & NON-LINEAR	
		VALIDATION	CROSS-VALIDATION	VALIDATION	CROSS-VALIDATION
100	REGRESSION	.77 (.04)	.76 (.04)	.77 (.03)	.74 (.04)
	NEURAL NET	.81 (.04)	.68 (.05)	.81 (.05)	.66 (.07)
1000	REGRESSION	.78 (.01)	.78 (.01)	.78 (.01)	.78 (.01)
	NEURAL NET	.77 (.01)	.76 (.01)	.77 (.01)	.77 (.01)

For each condition, the mean correlation is displayed followed by the standard deviation in parentheses. The predictor set used, made virtually no difference for these data. This result was expected because the additional non-linear terms should not be useful when predicting a linear function. The validation correlations in Table 2 tend to be near the 0.77 listed in Table 1 except for the value of 0.81 for neural networks on data sets of 100 cases. It should be noticed, however, that this prediction showed considerable shrinkage upon cross-validation. Finally, it can be seen that the regression based prediction consistently cross-validated better than the neural network predictions. Although, the neural network predictions

generalized better on data sets with 1000 cases than neural networks using 100 cases, the regression equation was statistically better than neural networks, on data sets of 1000 cases, both when linear predictors were used ($t = 8.82$, $p < .001$; $df = 9$) and when all predictors were used ($t = 4.80$, $p < .001$; $df = 9$).

Polynomial Function

The results of the analyses conducted to predict the outcome containing the second order polynomial are shown in Table 3. The correlations obtained with the linear predictors on data sets with 1000 cases shows that the regression technique had a validation correlation of 0.63 and the corresponding value for the neural network was 0.77. Comparing these numbers to the values in Table 1, it would appear that the regression equation is predicting the linear variance while the neural network is capturing both the linear and non-linear variance. Further, upon cross-validation, there was virtually no shrinkage for either prediction. On data sets with 100 cases, it appears that the regression equation slightly overfit the linear component while the neural network tended to overfit the data to a much greater extent. The extent of overfitting appears to have been reflected in the amount of shrinkage that occurred upon cross-validation. For the regression equation, the 0.65 fell to 0.62, while the 0.82 for the neural network fell to 0.68.

Table 3

Mean Correlation Between Actual and Predicted Criterion Scores Generated Using a Function with a Second Order Polynomial

		PREDICTOR SET			
NO. CASES	TECHNIQUE	LINEAR TERMS ONLY		LINEAR & NON-LINEAR	
		VALIDATION	CROSS-VALIDATION	VALIDATION	CROSS-VALIDATION
100	REGRESSION	.65 (.05)	.62 (.04)	.76 (.05)	.74 (.05)
	NEURAL NET	.82 (.04)	.68 (.06)	.82 (.05)	.67 (.07)
1000	REGRESSION	.63 (.02)	.63 (.02)	.78 (.01)	.77 (.01)
	NEURAL NET	.77 (.01)	.76 (.01)	.77 (.01)	.76 (.01)

When the non-linear terms were added to the predictor set, the correlations for the regression technique increased noticeably, while the neural network correlations remained virtually unchanged. Again this would suggest that the neural network was able to predict the non-linear variance without the additional terms, and when added, the non-linear terms were of no benefit to the network. Although, regression could not predict the interactive variance without the non-linear terms, when provided the non-linear information the regression technique was able to meet or exceed the performance of the neural network. When using data sets with 100 cases, the difference between 0.74 and 0.67 was statistically significant ($t = 6.13$, $p < .001$; $df = 29$) and under the 1000 case condition the difference was smaller, 0.77 versus 0.76, but still statistically significant ($t = 3.19$, $p < .05$; $df = 9$).

Interactive Function

The information from the analyses of the data sets with respect to the interactive function is shown in Table 4. As seen for the previous function, the validation correlations obtained on the 1000 case data sets are comparable to the values in Table 1. Again, it appears that given linear prediction terms, the regression fits only the linear variance while the neural network fits both the linear and non-linear. Further, when the non-linear terms are added to the predictor set, the regression solution performs well. In fact, cross-validation correlations of equations using linear and non-linear terms are better for the regression techniques than the neural network for both the 100 case data sets ($t = 7.26$, $p < .001$; $df = 29$) and 1000 case data sets ($t = 5.47$, $p < .001$; $df = 9$). Finally, it has been observed that for all three functions, the neural network tended to overfit the data sets with 100 cases and then suffer a considerable amount of shrinkage upon cross-validation.

Table 4

Mean Correlation Between Actual and Predicted Criterion Scores Generated Using a Function with an Interactive Component

		PREDICTOR SET			
NO. CASES	TECHNIQUE	LINEAR TERMS ONLY		LINEAR & NON-LINEAR	
		VALIDATION	CROSS-VALIDATION	VALIDATION	CROSS-VALIDATION
100	REGRESSION	.69 (.05)	.66 (.05)	.72 (.05)	.71 (.05)
	NEURAL NET	.78 (.05)	.60 (.07)	.78 (.08)	.60 (.10)
1000	REGRESSION	.68 (.02)	.68 (.02)	.73 (.01)	.73 (.01)
	NEURAL NET	.72 (.01)	.71 (.02)	.72 (.01)	.72 (.02)

Discussion

Analyses using only the linear predictors showed that the regression procedure was able to predict the linear component of the outcome variable while the neural network was able to fit both the linear and non-linear variance. So, when all the component predictors were related to the outcome in a linear fashion the neural network had no advantage over the regression. The advantage of using a neural network, it would seem, is that it will fit virtually any form of function that the predictors have with the criterion; the investigator does not have to model the relationship first.

This advantage, however, may be more apparent than real. First, the linear regression was able to predict the non-linear variable once the appropriate predictor transformations were made

and entered into the analysis. Therefore, regression can fit the same variance the neural network accounted for, but the investigator must provide information on the form of the relationship to the analytic technique. At this point, it should be noted that neural network must also be "tuned". The user must specify the number of layers to be used, the number of nodes at the hidden layers, the learning rate, and the activation function. It can be argued such "tuning" is the neural networkers way of providing information on the form of the relationship to the program.

To be able to predict the non-linear variance, the neural network is, in effect, generating an array of non-linear terms transparent to the user and entering those into the analysis. While this process can help if there is, in fact, a non-linear component to predict, it can also work against the user. In particular, if a large number of terms are being tested, then a large number of cases are needed to develop stable weight. Thus, when the number of cases is too small, the extra terms supplied by the neural network will fit erroneous variance. These errors will then work against the effort to predict the outcome in new data samples. As we saw, the neural network consistently overfit the data upon validation in the data sets with only 100 cases, and as a result, produced sub-optimal predictions upon cross-validation. Data sets with 1000 cases, however, provided enough information for the neural network to generate stable results.

For the researcher who is interested in developing diagnostic algorithms, these results provide some important guidance. First, one should be aware the predictor information may not have a linear relationship to the outcome being studied. To explore this possibility, one could use a neural network or multiple regression with non-linear terms included.

If the neural network is used, one should strive to use a large data set to develop the prediction parameters. It is also recommended that a cross-validation sample be used to test the result. A large degree of shrinkage between the validation and cross-validation predictions would suggest that the size of the validation sample was too small and the predictions could be improved.

If the number of cases for analysis is small, the investigat or should consider using regression, and if regression is used, the number of predictor terms used should be monitored. It is recommended that medical expertise be brought to bear before any analyses are conducted. This way any potential terms, linear and non-linear, can be identified and meaningful transformations can be generated. Again, it is recommended that parameters be developed on a validation sample and tested on a cross-validation sample. The prior modeling, however, should guard against a high degree of shrinkage in the predictive validities. In addition,

review of the regression weights will show which terms significantly predict the outcome.

This feature of regression programs; to identify those terms that are used to predict the criterion, is an important one which should be considered even if a neural network has successfully predicted an outcome. That is, the researcher may wish to use a hybrid approach where a neural network and a regression analysis are both conducted. If it appears that the neural network is accounting for more criterion variance, various transformations of the predictors could be tested. This way one could determine which variables are contributing to the prediction and the form of the predictive function would be known, while the accuracy of prediction would be maintained. Such information could be used to simplify predictive algorithms and contribute to knowledge about how specific diagnoses are made.

These analyses and results do not identify which technique is better, and that was not the intent of this exercise. Instead the relative strengths and weaknesses were demonstrated and discussed. The course of action that is taken by a researcher should include a consideration of the amount of data available for analysis, knowledge about the relationship between symptoms and diagnoses of interest, and the need to know how a specific diagnosis was achieved.

REFERENCES

Carras, B. G., Southerland, D. G. & Fisherkeller, K. D. (1989). PAIN; A Decision Support Program the Management of Acute Abdominal Pain, (Tech Report 1146). Groton, Connecticut: Naval Submarine Medical Research Laboratory.

de Dombal, F. T., Leaper, D. J., Staniland, J. R. McCann, A. P. & Horrocks, J.C. (1972). Computer-Aided Diagnosis of Acute Abdominal Pain, British Medical Journal, 2, 9-13.

Eberhart, R. C. & Dobbins, R. W. (Nov.1989). Neural Network Versus Bayesian Diagnosis of Appendicitis, Proc IEEE EMBS Annual Conf, Philadelphia, PA, 78-80.

Eberhart, R. C. & Dobbins, R. W. (1990). Neural Networks PC Tools: A Practical Guide, (Eds.) San Diego: Academic Press.

Kinney, E. L., Brafman, D. & Wright, R. H. (1988). An Expert System on the Diagnosis of Ascites, Computers and Biomedical Research, 21, 169-173.

Ledley, R. S. & Ted, L. B. (1959). Reasoning Foundation of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid In Understanding of How Physicians Reason, Science, 130, 9-22.

Newacheck, J. S. (1990). Computer Aided Ocular Assessment - Programmers Manual, Groton, Connecticut, Naval Submarine Medical Research Laboratory.

Stetson, D. M., Eberhardt, R. C., Dobbins, R. W., Pugh, W. M. & Gino, A. (1990). Structured Specification of a Computer Assisted Medical Diagnostic System, Computer Assisted Medical Diagnostic System, Chapel Hill, N. Carolina.

Titterington, D. M., Murray, G. D., Murray, L. S. & Spgiegelhalter, D. J. (1981). Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients, J. Royal Statistical Society, 144, 145-161.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	10-91	Interim Jul-Sep 91

4. TITLE AND SUBTITLE
(U) A Comparison of Multiple Regression and a Neural Network for Predicting a Medical Diagnosis

5. FUNDING NUMBERS
Program Element: 63706N
Work Unit Number:
M0095.005-6051

6. AUTHOR(S)
Pugh, William M., Ryman, David H.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

8. PERFORMING ORGANIZATION
Report No. 91-33

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Naval Medical Research and Development Command
National Naval Medical Center
Building 1, Tower 2
Bethesda, MD 20889

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NHRC Tech Report

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Regression and neural network prediction methods were compared using artificial data generated to simulate three types of predictor-criterion relationships: linear, polynomial, and interactive. Analyses of linear data indicated that both methods were comparable on large data sets.

On small data sets the neural network tended to overfit the initial data and thus did not generalize as well as the regression equation. Analysis of data with a non-linear component demonstrated the ability of the neural network to fit either a polynomial or interactive term without the user having to model such terms. However, when these effects were modeled, the regression equation permored well. The implications of these results for the development of predictive algorithms were discussed.

14. SUBJECT TERMS
algorithms
predictive algorithms
neural network
regression equations
diagnostic algorithms

15. NUMBER OF PAGES
22

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited